Understanding the Effect of Baseline Modeling Implementation Choices on Analysis of Demand Response Performance

Nathan Addy, Sila Kiliccote
Lawrence Berkeley National Laboratory
Johanna Mathieu, Duncan S. Callaway
University of California, Berkeley

June 2012

DISCLAIMER

Understanding the Effect of Baseline Modeling Implementation Choices on Analysis of Demand Response Performance

Nathan Addy
Environmental Energy Technologies Division
Lawrence Berkeley National Lab
Berkeley, California, 94720
Email: naddy@lbl.gov

Johanna L. Mathieu
Department of Mechanical Engineering
University of California, Berkeley
Berkeley, California, 94720
Email: jmathieu@berkeley.edu

Sila Kiliccote
Environmental Energy Technologies Division
Lawrence Berkeley National Lab
Berkeley, California, 94720
Email: skiliccote@lbl.gov

Duncan S. Callaway
Energy and Resources Group
University of California, Berkeley
Berkeley, California, 94720
Email: dcal@berkeley.edu

June 15, 2012

Abstract

Accurate evaluation of the performance of buildings participating in Demand Response (DR) programs is critical to the adoption and improvement of these programs. Typically, we calculate load sheds during DR events by comparing observed electric demand against counterfactual predictions made using statistical baseline models. Many baseline models exist and these models can produce different shed calculations. Moreover, modelers implementing the same baseline model can make different modeling implementation choices, which may affect shed estimates. In this work, using real data, we analyze the effect of different modeling implementation choices on shed predictions. We focused on five issues: weather data source, resolution of data, methods for determining when buildings are occupied, methods for aligning building data with temperature data, and methods for power outage filtering. Results indicate sensitivity to the weather data source and data filtration methods as well as an immediate potential for automation of methods to choose building occupied modes.

1 Introduction

Buildings are becoming increasingly important components of power system operations. With continuing Smart Grid development, there is a potential for electric loads to become dynamic resources that are as equally controllable as electricity generators [1]. In traditional demand response (DR) programs, system operators and utilities can achieve system-wide demand reductions by providing financial incentives for buildings to reduce their demand during time periods when the grid is stressed. In dynamic pricing programs, operators incentivize behavior by increasing electricity prices during periods of grid stress, encouraging building operators to shed or shift load to an off-peak time.

DR programs are evaluated by their impact and cost effectiveness. Accurate estimations can significantly change these values. Therefore, a key to the success of DR programs is the accurate estimation of demand sheds achieved by program participants. These estimations are typically made with baseline models that estimate what building load would have

been if a DR event had not been called. These estimates are compared with actual measurements of building data to estimate the size of sheds, or load curtailments. These baseline models are used for a variety of tasks including Measurement and Verification (M&V), improving DR program design and operation, and, in some cases, settling business transactions surrounding DR events.

There are many examples of baseline models in the energy efficiency literature [2, 3, 4, 5, 6, 7, 8] and the DR literature [9, 10, 11, 12]. Some of these studies compare estimates produced by different baseline models. However, shed estimates *from the same model* can differ if the model is implemented by two different building modelers. This is because specific algorithm implementation choices can effect model results. For example, different approaches to interpreting and filtering bad data, different methods for calculating model parameters, and different sources of model inputs can all affect final baseline predictions. This issue is of importance because as individual modeling frameworks become more widely used, the effects of implementation differences could cause differences in interpretation of DR performance. Therefore, it is important to understand which sorts of differences have the most effect on results.

In this work, we use a linear regression model relating time-of-week, outdoor air temperature, and whether or not the building is in an occupied mode to building demand as described in [13]. We reimplemented this model on a new platform and describe lessons learned through validation. Then we look at five variations on choices made in the original implementation: (1) different sources of weather data, comparing the National Climactic Data Center data used in the original analysis, which is heavily curated but at a lower time-resolution and usually measured further from the sites, to Weather Underground data, which is less curated, higher resolution, and measured closer to the sites; (2) different resolution of building data, comparing the predicted sheds using 15-, 30-, and 60- minute resolved data; (3) different approaches for determining whether the building was in an occupied or unoccupied mode, with the transition times either estimated manually/visually or with an algorithm that automatically calculates these transition times based on a heuristic; (4) different methods for aligning building electric demand data with temperature data; and (5) the choice of a model parameter that determines sensititivity to identifying power outages in the building data.

The rest of this paper is organized as follows. Section 2 describes the data we used in this analysis. In Section 3, we describe the baseline model as well as its validation against the original implementation. Section 4 discusses the details of the modeling variations we examined in this work. Finally, in Sections 5 and 6, we discuss and conclude the work.

2 Data

For the base analysis, we use 15-minute interval whole building electric load data from 38 large commercial buildings and industrial facilities in the Pacific Gas and Electric Company's (PG&E) Automated Critical Peak Pricing (CPP) Program between 2007 and 2009. In the CPP program, on up to 12 days per year electricity prices were raised to three times the peak price between 12 pm and 3 pm in a moderate price period, and raised to five times peak price in a high priced period between 3 pm and 6 pm. These 'DR events' were announced day-ahead.

In the base analysis, we used weather data obtained from the National Climactic Data Center (NCDC) [14], a division of the National Oceanic and Atmospheric Administration (NOAA). Hourly outdoor air temperature (OAT) was downloaded for the nearest facility to each site. Linear interpolation was used to approximate OAT at each 15 minute interval. Weather data were removed for times when the interval between interpolants was greater than six hours. In some cases, where exceptionally large amounts of data were missing, temperature files were patched using OAT from the second closest NOAA weather station.

To investigate the effect of weather data source on shed estimates, we obtained OAT data from Weather Underground [15]. Weather Underground is a private site that collects data from Personal Weather Stations (PWS) operated by private individuals and or-

ganizations. Stations undergo a one-time calibration, but are generally not guaranteed to be monitored by meteorological experts. However, time resolution tends to be much higher and and for many locations, typically occurring in high density areas, multiple measurements of OAT can typically be found closer than those used by NOAA. Additionally, in general, these weather stations had better up-time and so the data were less spotty than the NOAA data.

Because Weather Underground data are collected from a variety of sources, specifics of data format may vary. Each weather station collects a variety of measurements; however OAT is measured at essentially all stations. Weather temperature data measurement rate is dependent on the specific weather station being queried, however 5- or 15-minute weather data are typical.

3 Baseline Model

In this section, we briefly describe the baseline model in [13] that we used in this analysis. We build separate models for each facility in each year (facility-year) since buildings change year to year. As in [13], for each facility-year, we use five months (May 1 – Sept 30) of load and temperature data for each model.

The model assumes demand is a function of time of week, and assigns a regression coefficient α_i to each 15-minute interval from Monday through Friday, t_i where $i = 1 \cdots 480$. The model also assumes that demand, when the building is occupied, is a piecewise linear and continuous function of OAT. To model this effect, each observed temperature is divided into six temperature components, $T_{c,j}$ with $j = 1 \cdots 6$, associated with six equal sized temperature bins. A regression coefficient β_j, is assigned to each bin. $T_{c,j}$ is computed with the following algorithm:

1. Let B_K for $k = 1 \cdots 5$ be the interior bounds of the temperature intervals.

2. If $T > B_1$ then $T_{c,1} = B_1$. Otherwise, $T_{c,1} = T$ and $T_{c,m} = 0$ for $m = 2 \cdots 6$ and algorithm ended.

3. For $n = 2 \cdots 4$, if $T > B_n$ then $T_{c,n} = B_n - B_{n-1}$. Otherwise, $T_{c,n} = T - B_{n-1}$ and $T_{c,m} = 0$ for $m = (n+1) \cdots 6$ and algorithm is ended.

4. If $T > B_5$ then $T_{c,5} = B_5 - B_4$ and $T_{c,6} = T - B_5$.

Estimated occupied mode demand, $\hat{D_o}$, is calculated as:

$$\hat{D_o}(t_i, T(t_i)) = \alpha_i + \sum_{j=1}^{6} \beta_j T_{c,j}(t_i) \tag{1}$$

A different temperature effect is modeled during unoccupied hours of the building. During building unoccupied mode temperature is modeled using only one regression coefficient, β_u which is multiplied by outdoor temperature T. Estimated unoccupied mode demand, $\hat{D_u}$, is calculated as:

$$\hat{D_u}(t_i, T(t_i)) = \alpha_i + \beta_u T(t_i) \tag{2}$$

Ordinary Least Squares (OLS) is used to estimate the parameters α, β, and β_u. It produces unbiased regression parameters. Because of autocorrelation and heteroscedasticity, we do not use the standard errors associated with the regression parameters.

The general procedure to build the model is as follows. We take building demand data and filter out weekends, holidays, and days on which buildings participated in DR events. We filter for power outages by looking for days where minimum power consumption was less than 50% of the average minimum daily power consumption during the summer. For any day that falls below the threshold, we flag the entire day's data as having a power outage and remove it from the analysis.

Next, we take the outdoor temperature data and linearly interpolate it to 15 minutes prior to the time stamp on the building data. We do this because for 15-minute interval building data, each measurement represents the mean demand by that building over the previous 15 minute interval. After interpolation, we finally filter out all times where the temperature values were computed using interpolants greater than 6 hours apart. This represents the final, cleaned set of data used to build the model.

Next, since buildings use electricity differently when they are occupied than when they are unoccupied, we determine transitions between unoccupied

and occupied mode (usually in the morning) and occupied model and unoccupied mode (usually in the evening). In the original analysis these start and end occupied mode transition times were determined through visual inspection of the load shape data. An algorithm for doing this is presented in Section 4.

The observed temperature range is divided into 6 regions according to the procedure described above, so that for a given occupied temperature we can decompose it into the contributions for each range. For example, if the minimum observed temperature is 40°F (4.4°C) and the maximum is 100°F (37.7°C), then the minimum bin starts at 40°F and each of the 6 bins has width 10°F. If we decompose a temperature of 65°F, then $T_{c,1} = 50$°F, $T_{c,2} = 10$°F, $T_{c,3} = 5°$, and the remaining temperature components are 0°F.

The regression equations (Eqns. 1 and 2) can be written in matrix form:

$$y = Ax + \epsilon \tag{3}$$

where x is the parameter vector, y is the output vector (electric demand), ϵ is the error, and A is composed of time of week indicator variables, occupied mode temperatures components, and unoccupied mode temperatures. For each building time measurement and temperature that has not been filtered away, we generate a 487-column row vector. The first 480 columns correspond to the time of week indicator variables which are 0 or 1 depending on the 15-minute interval of the Monday through Friday work week; columns 481-486 correspond to the occupied mode temperature components; and column 487 is the unoccupied mode temperature. After building the A matrix by generating each of its rows, we solve for the parameter vector x using an OLS estimator:

$$\hat{x} = (A^T A)^{-1} A^T y \tag{4}$$

In practice this is calculated using the implementation of the software package being used to implement the model.

To make a prediction for a given time-of-week and temperature, we generate the corresponding 487 column row vector, v, and predict:

$$y_{predict} = v \cdot x \tag{5}$$

To estimate average demand shed over a period, we make a prediction for each of the relevant times of week, subtract the observed demand, and take the mean.

Model Validation

The model described in the previous section was originally implemented in MATLAB. To do the analysis described in this paper, we reimplemented the model in Python and validated this implementation against the results of the original implementation. While our primary focus was simply to verify that the new implementation performed correctly, the validation process also helped us gain a sense for the variety of important modeling implementation choices that modelers face, and the implications of those choices. These choices ranged from the technical, such as rounding choices that encouraged floating point disagreements between estimates made on different computer systems, to the more pragmatic, such as a strong influence of the effect of thresholds to filtering algorithms whose differences caused the model to be built on different subsets of data. This experience helped us pick the set of modeling choices to investigate, described in the next section. Additionally, it left us with a number of lessons learned, described in Section 5.

We validated the Python implementation via a two stage process. We first looked at five facility-years worth of data in detail, performing an end-to-end comparison between the two implementations, identifying and classifying as many discrepancies as could be found. After the detailed validation, a statistical analysis was completed on the full set of data with the purpose of comparing the population shed estimates from one implementation to the other. At this point, outliers were identified visually and where appropriate issues were tracked down until the authors were satisfied the two implementations behaved substantially identically.

Figure 1 shows the comparison between the estimates of the first analysis and the second analysis. Each point represents a comparison between a DR shed (one for the moderate price period and one for the high price period for each facility-year) calculated

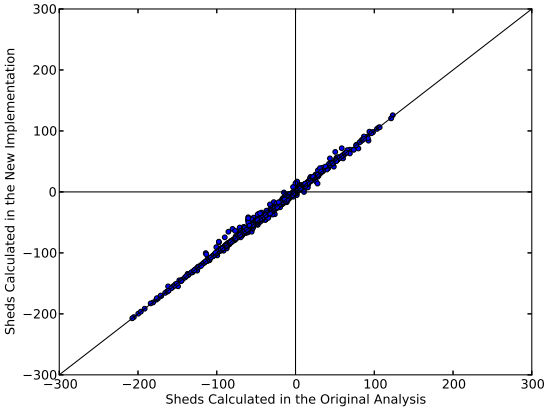

Figure 1: FINAL RESULT OF VALIDATION.

using the first compared with the second implementation.

The remaining discrepancies were caused by a variety of minor factors. Because of a choice to round interpolated OAT data, temperature values used by identical algorithms on two different machines occasionally differed (details described in Section 5). Additionally, the boundaries on the weather filter were different in the two implementations, resulting in a small number of weather points being used by one implementation and not the other. Finally, a slightly different power outage filtering routine was used in the second implementation; sites on which the two filters differed were removed from the analysis, so as not to bias results. Ultimately, 49 facility-years worth of data (out of the original 87 facility-years of data) were used to perform the analyses in the subsequent sections. In total, each analysis includes 1176 shed estimates.

4 Modeling Choices Investigated

The goal of our analysis was to gain a general sense of the relative importance of different potential modeling implementation choices focusing on five types of choices: choice of weather data, choice of building load data resolution, choice of method to determine occupied/unoccupied mode transition times, choice of alignment of OAT data with building demand data, and choice of power outage filter. This analysis does not attempt to be comprehensive for each modeling choice, but instead seeks to test plausible real world choices, to get a better sense of what the contentious choices might be and where future efforts in model building might be directed.

Therefore, for each type of modeling implementation choice investigated, we looked at two or three different choices that could be made and the effect of those choices on the resulting analysis when compared with the base analysis.

For each choice, we calculate the sheds generated using the validated baseline model (producing the 'base analysis') and then generate the sheds using the model with variations ('variant analysis'). We calculate a variety of statistics on these predictions to gain a sense of the effect of the two choices on shed prediction. For each shed predicted in each analysis, we calculate the mean. Additionally, we look at the differences between the base and variant sheds, calculated as $(shed_{variant} - shed_{base})$, and report the mean and variance between the differences. Assuming the differences are unbiased between the original and mean analysis, we expect the mean difference between sheds to be near zero. We calculate the absolute mean difference for each shed, calculating $|shed_{variant} - shed_{base}|$, and take the mean and variance of these values. We also calculate the relative difference in sheds, as:

$$\left| \frac{shed_{variant} - shed_{base}}{shed_{base}} \right|$$

We calculate the mean and variance of the relative differences as well. Statistics are summarized in Tab. 1.

Weather Data Source

To get a sense for the effect of choice of weather sources, we compared the base analysis which uses NOAA OAT data against a variant analysis which used Weather Underground OAT.

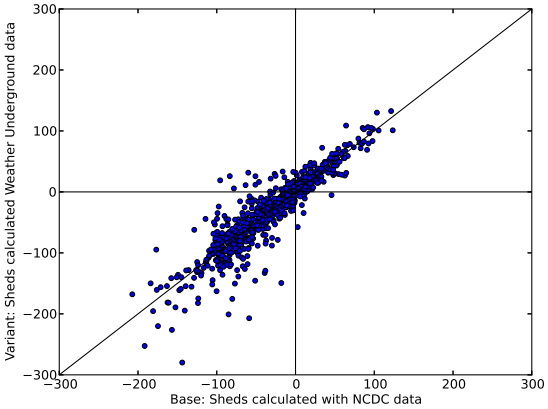

Figure 2: COMPARISON OF BASE ANALYSIS TO VARIANT ANALYSIS USING WEATHER UNDERGROUND DATA.

For each facility, we used its zip code to look up the closest weather stations using the Weather Underground website and downloaded data from the two sites with relevant data. When these two stations differed in distance to the zip code by more than 50%, we used data from the closest site directly. Otherwise, we averaged the two data streams when both were available and directly used data from one of the two sites when only one was available. For several sites, we were not able to easily obtain good weather data from Weather Underground. We removed these sites from both the base and modified analysis for this specific comparison in order to generate good statistics.

The results of the comparison between NOAA and Weather Underground are shown in Fig. 2. Shed statistics are summarized in Tab. 1.

Building Data Resolution

Building models may be built using various resolutions of load and weather data. This choice may be made either through a choice of sensor configuration or it may be made at model building time, by downsampling.

To get a sense of the effect of building demand data

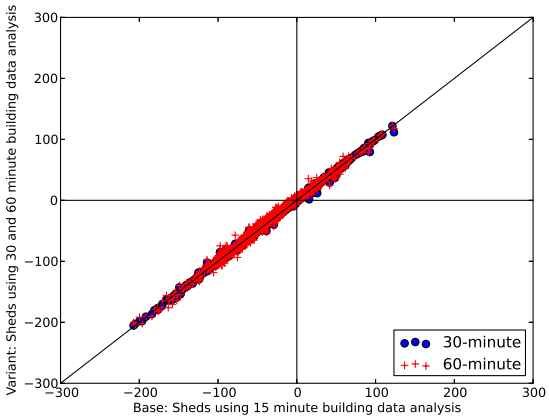

Figure 3: COMPARISON OF BASE ANALYSIS TO VARIANT ANALYSES USING 30- AND 60-MINUTE INTERVAL BUILDING DEMAND DATA.

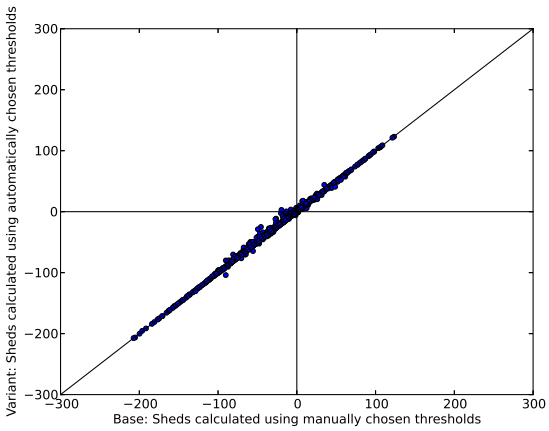

Figure 4: COMPARISON OF BASE ANALYSIS TO VARIANT ANALYSIS USING AN AUTOMATED OCCUPIED MODE DETECTION ALGORITHM.

Table 1: COMPARISON OF STATISTICS BETWEEN BASE AND VARIANT ANALYSES.

	Weather Underground data	30-minute interval data	60-minute interval data	Automated occupied mode detection	Shift data alignment	No Power Outage Filter	Sensitive Power Outage Filter
Mean Shed using Base Analysis (kW)	-32.4	-34.5	-34.5	-34.5	-34.5	-34.5	-34.5
Mean Shed using Variant Analysis (kW)	-32.0	-34.4	-33.7	-34.0	-33.1	-30.3	2.5
Mean Difference Between Sheds (kW)	-0.4	-0.1	-0.9	-0.6	-1.4	-4.3	9.3
Mean Absolute Shed Differenc (kW)	14.6	1.3	3.3	1.0	3.2	4.6	0.7
Mean Relative Shed Difference (kW)	1.4	0.2	0.4	0.1	0.3	0.9	344.2
Variability Between Sheds (kW^2)	546.1	4.9	21.1	5.7	21.1	904.9	263.9
Variability of Absolute Shed Diff. (kW^2)	332.0	3.1	11.1	5.1	12.6	901.7	263.9
Variability of Relative Shed Diff. (kW^2)	546.1	4.9	21.1	5.7	21.1	904.9	344.2

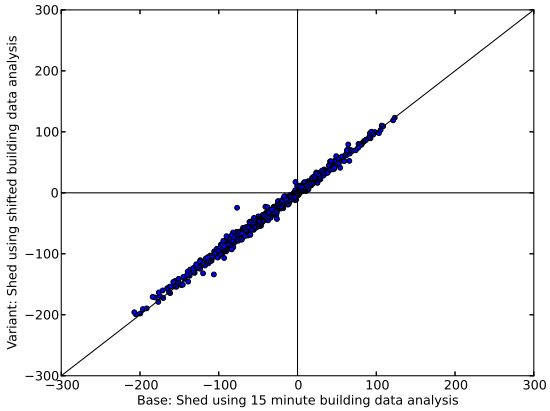

Figure 5: COMPARISON OF BASE ANALYSIS TO VARIANT ANALYSIS USING DIFFERENT ALIGNMENT OF OAT AND DEMAND DATA.

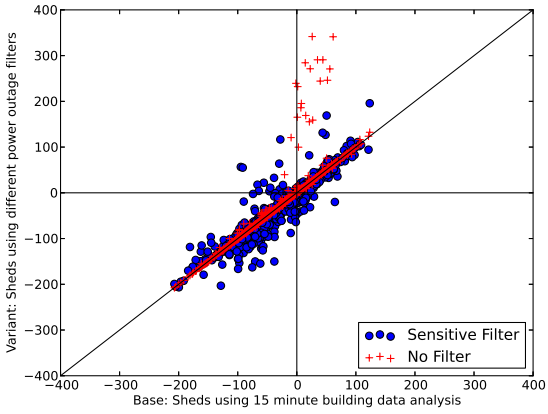

Figure 6: COMPARISON OF BASE ANALYSIS TO VARIANT ANALYSES USING DIFFERENT THRESHOLDS FOR FILTERING POWER OUTAGE DAYS.

resolution on shed estimates, we took the original 15-minute interval data and decreased the resolution to 30- and 60-minute interval data.

For each time series used in this analysis, we calculated 30- and 60-minute resolution data by finding each 30- or 60-minute period worth of data and taking the mean of those values. Care was taken to ensure the intervals were day-aligned, meaning that the first interval of the day always represented demand during the interval from midnight to either 00:30 or 01:00.

If one or more data points were missing, we skipped over that point in the algorithm, e.g., if a 30-minute interval only had zero or one data point, or the 60-minute interval contained zero, one, two, or three data points, they were skipped over and not included during the analysis. This had a minimal effect since sites typically had fewer than 10 hours of data thrown out during this process.

Once completed, we ran the analysis to generate shed predictions the 30- and 60-minute resolution data. The results of the 15- vs. 30-minute analysis and the 15- vs -60 analyses are plotted in Fig. 3 and statistics are summarized in Tab. 1.

Occupied/Unoccupied Mode Transitions

In the model, the occupied mode of the building is an implicit variable because temperature effects are modeled differently depending on the mode, according to either Eqn. 1 or 2.

In the original algorithm, occupied periods were determined manually by visual inspection of the data. This process has all the advantages and disadvantages of having a human in the loop. This can be compared with an algorithmic approach to determining occupied and unoccupied period transition times.

To gain a sense of the effect of this modeling choice on the outcome to the analysis, we created a basic algorithm to automatically detect the occupied mode start and end times. The algorithm looks at each day of the analysis and calculates the 2.5th and 97.5th percentiles, called $D_{2.5}$ and $D_{97.5}$ of load. These percentiles were chosen based on work in [16] which used them in order to minimize the effect of demand outliers skewing the analysis. For each day, the transition time from unoccupied to occupied mode ('start time'), typically in the morning, was determined by calculating the first time the building transitioned above $0.1 \times (D_{97.5} - D_{2.5}) + D_{2.5}$. The transition from occupied to unoccupied mode ('end time') was calculated as the final time during the day the building transitioned below this threshold. The mean of each facility-year's start-times and end-times were used to determine when each building was in occupied and unoccupied modes for the purposes of the model.

The results comparing the base analysis to the variant analysis with an automated occupied mode detection algorithm is shown in Fig. 4. Shed statistics are summarized in Tab. 1.

Alignment of OAT Data with Building Demand Data

To build the model, each demand measurement is associated with an OAT measurement. In the original work, OATs were assigned to the beginning of the interval over which the building demand measurements were taken. For example, with 15-minute interval data, the demand measurement at 3:00pm was assigned an OAT at 2:45pm. We tested the effect of assigning OAT data based on the end of the building demand interval measurement, i.e., matching 3pm to 3pm, a simpler choice. The results are shown in Fig. 5. Shed statistics are summarized in Tab. 1.

Sensitivity of Power Outage Filter

In the model, a day is flagged as being a power outage day and filtered if its daily minimum demand falls below some threshold percentage of the mean daily minimum demand for the dataset. In the original work, this threshold was set to 50%. To test the effects of permitting borderline data to enter the analysis, we ran the analysis using no power filter whatsoever. We also tested the effects of running with a stronger filter that flags days with a measurement below a 75% of average daily minimum threshold. The results are in Fig. 6 and the shed statistics are reported in Tab. 1.

5 Discussion

Effect of Modeling Choices on Shed Estimation

We find a strong sensitivity on shed estimation to the source of weather data. Especially in areas with strong micro-climate effects, it may be worth spending time and investment in obtaining good weather data, including installing additional sensors as needed. Where this is not possible, it may be worth acquiring multiple sources of weather data and running the analysis with on different weather data, to gain a sense of the differences in shed estimates. It is clear that the impact of real work weather data on baseline model predictions is a topic needing further investigation.

Investigating the choice of building data resolution, we found a moderate sensitivity towards using coarser 60-minute building interval data compared with 15-minute data, and a relatively slight effect when using 30-minute over 15-minute data. For this data set, it appears that discrepancies do not increase substantially as data set is coarsened, at least over this range. This suggests that if there are compelling reasons to do so, it is likely acceptable to use a coarser grained building data set.

We also determine high levels of agreement between manually determining building occupancy mode thresholds and automatically detecting it using a very simple algorithm. While we make no claim that this algorithm is optimal for this task, even a basic approach agrees very well with manually choosing the times. From this, we conclude that automatic detection is beneficial, providing very similar performance while eliminating the need for a human in-the-loop, speeding up processing time.

We find there are only minor effects associated with demand and OAT data alignment, at least for plus/minus 15-minutes. Although offsetting OATs by 15 minutes against the building demand data may be more accurate, this subtle complexity could be a potential source of invisible disagreements between tools in the future. This result suggests that it may be possible to opt for a simpler approach without noticeably affecting results, especially given the apparent relative sensitivity to the weather data.

We find large differences caused by filtering power outages. Each day of marginal data has an outsized effect on the ultimate estimation of model parameters, and therefore the choice of this parameter matters greatly to the overall analysis. This has multiple implications. First, it is likely worth investing resources into developing good algorithms to detect power outages. It may also be worth obtaining information on power outages directly, rather that estimating them, as disagreements in this area may lead to outsized effects. Further investigation is warranted.

Lessons Learned through Algorithm Validation

During the course of validating the model, we learned several practical lessons that may be of use to other implementers. Depending on the situation, this could either save development time for future research or applications, or it could even help prevent discrepancies in code that gets used in the field, as in many real-world examples replicating someone else's analysis could be difficult, as much data is protected from being shared by non-disclosure agreements (NDAs).

Generally, we found this process to be incredibly helpful in better understanding performance of the model. Several modeling implementation issues were discovered during this process.

The first was the discovery of a choice to round interpolated temperatures. This choice was significant, because it caused discrepancies due to numerical floating point calculations in the rounding process across different machines. One machine would linearly interpolate a value ending in .5 and would round up; another machine would calculate a value of .49999... and round down. These individual discrepancies appeared to occur arbitrarily. While we were not able to discern a significant difference in model prediction caused by this choice, from a software development perspective, this choice confers little added benefit and increases the probability of discrepancies. For instance, comparing intermediate results when such a choice has been made could make things more difficult than necessary.

The second discovery was the significance of the algorithms used to filter out bad building and temperature data. Many of the discrepancies we tracked had to do with specifics as to how these filters were applied. In most cases, we were surprised by the large effect of these filtering parameters. While we did not investigate these choices directly in this work, we suggest testing various settings against one another, to characterize the effect of including or not including marginal data on analyses.

Finally, during the validation, we also discovered that there were several unexpected pitfalls caused by external software purporting to do the same thing but actually not. For instance, between the MATLAB and Python computer environments, the default variance calculation had a different interpretation as either sample or population variance. Implementation of the two *regress* functions treated NaNs very differently. Both these caused initial discrepancies. Especially where detailed analyses are not available with which to validate one tool against another, we recommend testing external computer routines against known inputs, to ensure that the semantics are as expected.

6 Conclusion

We have investigated the sensitivity DR performance results to different modeling implementation choices. We find that shed estimates are sensitive to weather data and therefore acquisition of good weather data should be a key focus of any baseline analysis. This may have implications for whether it is appropriate to install local weather monitoring or not. In areas where local climate varies noticeably, even around the shell of a building, it may be worthwhile to spend added resources finding optimal weather data.

We find shed estimates are not sensitive to building demand resolution. Depending on the application, it may be an acceptable trade off to use lower resolved building data, at least up to hourly, as other factors warrant.

We find that automated approaches for determining building occupied mode work essentially as well as manual approaches for this data set. Ultimately, for large data sets, automated approaches are necessary in order to increase the throughput of these analyses. We also find that short time-scale (plus/minus 15-minute) alignment of OAT and demand data has a relatively minor effect on model prediction. It may be advisable to standardize on simpler approaches. Finally, we find that the choices surrounding power outage filtration and, by extension, other data filtration schemes that flag and remove marginal data have a large influence on predictions. It may be worth expending extra effort to ensure data quality so as to avoid having to heuristically estimate such important information.

We also suggest defining a robust validation method on these algorithms, especially due to the fact that certain comparisons may be difficult or impossible given data access requirements. Therefore, it is of particular importance to understand the subtle differences between implementations of common algorithms such as the variance and regress implementations. If is is not possible to fully validate in an end-to-end manner, we determine that at minimum software should be tested on a common data set to ensure a common language between software tools.

Acknowledgments

The authors would like to thank Pacific Gas and Electric Company for providing electric load data and funding this work through their Demand Response and Emerging Technologies Programs. This work was sponsored in part by the Demand Response Research Center which is funded by the California Energy Commission (Energy Commission), Public Interest Energy Research (PIER) Program, under Work for Others Contract No. 500-03-026 and by the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

References

[1] Callaway, D., and Hiskens, I., 2011. "Achieving controllability of electric loads". *Proceedings of the IEEE,* ***99***(1), pp. 184–199.

[2] Fels, M., 1986. "PRISM: an introduction". *Energy and Buildings,* ***9***(1-2), pp. 5–18.

[3] Katipamula, S., Reddy, T., and Claridge, D., 1998. "Multivariate regression modeling". *Journal of Solar Energy Engineering,* ***120***, p. 177.

[4] Kissock, J., Reddy, T., and Claridge, D., 1998. "Ambient-temperature regression analysis for estimating retrofit savings in commercial buildings". *Journal of Solar Energy Engineering,* ***120***, p. 168.

[5] Kissock, J., and Eger, C., 2008. "Measuring industrial energy savings". *Applied Energy,* ***85***(5), pp. 347–361.

[6] Goldberg, M., 2000. "A geometric approach to nondifferentiable regression models as related to methods for assessing residential energy conservation". PhD thesis, Princeton University, Princeton, NJ.

[7] Ruch, D., Kissock, J., and Reddy, T., 1999. "Prediction uncertainty of linear building energy use models with autocorrelated residuals". *Journal of solar energy engineering,* ***121***, p. 63.

[8] Yang, J., Rivard, H., and Zmeureanu, R., 2005. "On-line building energy prediction using adaptive artificial neural networks". *Energy and buildings,* ***37***(12), pp. 1250–1259.

[9] Coughlin, K., Piette, M., Goldman, C., and Kiliccote, S., 2009. "Statistical analysis of baseline load models for non-residential buildings". *Energy and Buildings,* ***41***(4), Apr., pp. 374–381.

[10] Goldberg, M., and Agnew, G., 2003. Protocol development for demand response calculation–findings and recommendations. Tech. Rep. CEC 400-02-017F, California Energy Commission (KEMA-XENERGY).

[11] Kozikowski, D., Breidenbaugh, A., and Potter, M., 2006. The demand response baseline, v.1.75. Tech. rep., EnerNOC OPS Publication.

[12] Wi, Y.-M., Kim, J.-H., Joo, S.-K., Park, J.-B., and Oh, J.-C., 2009. "Customer baseline load (cbl) calculation using exponential smoothing model with weather adjustment". In Transmission Distribution Conference Exposition: Asia and Pacific, 2009, pp. 1 –4.

[13] Mathieu, J., Price, P., Kiliccote, S., and Piette, M., 2011. "Quantifying changes in building electricity use, with application to demand response". *IEEE Transactions on Smart Grid,* ***2***(3), pp. 507–518.

[14] NOAA, 2009. National climatic data center. National Oceanic and Atmospheric Administration.

[15] Wunderground, 2012. Weather underground, http://weatherunderground.com.

[16] Price, P., 2010. Methods for quantifying electric load shape and its variability. Tech. Rep. LBNL-3713E, Lawrence Berkeley National Laboratory.